4

BAYESIAN MODEL SELECTION IN SOCIAL RESEARCH

*Adrian E. Raftery**

It is argued that *P*-values and the tests based upon them give unsatisfactory results, especially in large samples. It is shown that, in regression, when there are many candidate independent variables, standard variable selection procedures can give very misleading results. Also, by selecting a single model, they ignore model uncertainty and so underestimate the uncertainty about quantities of interest. The Bayesian approach to hypothesis testing, model selection, and accounting for model uncertainty is presented. Implementing this is straightforward through the use of the simple and accurate BIC approximation, and it can be done using the output from standard software. Specific results are presented for most of the types of model commonly used in sociology. It is shown that this approach overcomes the difficulties with *P*-values and standard model selection procedures based on them. It also allows easy comparison of nonnested models, and permits the quantification of the evidence *for* a null hypothesis of interest, such as a convergence theory or a hypothesis about societal norms.

This research was supported by NIH grant no. 5R01HD26330. I would like to thank Robert Hauser, Michael Hout, Steven Lewis, Scott Long, Diane Lye, Peter Marsden, Bruce Western, Yu Xie, and two anonymous reviewers for detailed comments on an earlier version. I am also grateful to Clem Brooks, Sir David Cox, Tom DiPrete, John Goldthorpe, David Grusky, Jennifer Hoeting, Robert Kass, David Madigan, Michael Sobel, and Chris Volinsky for helpful discussions and correspondence. I may be contacted by email at raftery@stat.washington.edu.

*University of Washington

1. INTRODUCTION

P-values and significance tests based on them have traditionally been used for statistical inference in the social sciences. In the past 15 years, however, some quantitative sociologists have been attaching less importance to *P*-values because of practical difficulties and counterintuitive results.

These difficulties are most apparent with large samples, where *P*-values tend to indicate rejection of the null hypothesis even when the null model seems reasonable theoretically and inspection of the data fails to reveal any striking discrepancies with it. Because much sociological research is based on survey data, often with thousands of cases, sociologists frequently come up against this problem. In the early 1980s, some sociologists dealt with this problem by ignoring the results of *P*-value-based tests when they seemed counterintuitive and by basing model selection instead on theoretical considerations and informal assessment of discrepancies between model and data (e.g., Fienberg and Mason 1979; Hout 1983, 1984; Grusky and Hauser 1984).

Then, in 1986, Bayesian hypothesis testing was brought to the attention of sociologists, particularly using the simple BIC approximation (Schwarz 1978; Raftery 1986*b*). This seemed to lead to intuitively reasonable results when *P*-values did not, and retrospectively validated some of the "common sense" decisions made in spite of *P*-values by the researchers mentioned above. As a result, BIC has become quite popular for model selection in sociology, particularly in log-linear and other models for categorical data.

Two other difficulties with the use of *P*-values for model selection are also prevalent in sociology, although they are less obvious. They arise when many statistical models are implicitly considered in the earlier stages of a data analysis. This happens when many possible control variables are measured, and one must decide which ones to include in the final model. Often this choice is made using a strategy that involves a collection or sequence of *P*-value-based significance tests, either informally by screening the *t*-values in the full model with all variables included and removing the least significant ones, or more formally by stepwise regression and its variants.

The first difficulty is that *P*-values based on a model selected from among a large set of possibilities no longer have the same interpretation that they did when only two models were ever considered (Miller 1984, 1990). Indeed, the use of *P*-values following

model selection can be dramatically misleading (Freedman 1983; Freedman, Navidi, and Peters 1988).

The second difficulty is that several different models may all seem reasonable given the data but nevertheless lead to different conclusions about questions of interest. This can happen even when the dataset is moderately large, and striking examples have been observed in educational stratification (Kass and Raftery 1995) and epidemiology (Raftery 1993*b*). In this situation, the standard approach of selecting a single model and basing inference on it underestimates uncertainty about quantities of interest because it ignores uncertainty about model form.

The Bayesian approach to model selection and accounting for model uncertainty overcomes these difficulties. It was first used in sociology in 1986 purely as a model selection criterion, and since then it has been widely applied. Here my aim is to give the rationale behind it, to show how it avoids the problems that plague *P*-values, to explain how it can be used to account for model uncertainty as well as to select a single "best" model, and to give some guidelines on its practical implementation for specific model classes.

In Section 2 I review some of the practical difficulties with *P*-values in empirical research and give examples. In Section 3 I give the basic ideas of Bayesian hypothesis testing and Bayes factors. In Section 4 I derive the BIC approximation and equivalent expressions useful for specific models used in social research. I discuss the interpretation of BIC and why it sometimes leads to different conclusions than *P*-values. In particular, BIC tends to favor simpler models and null hypotheses more than do *P*-values in large data sets. In Section 5 I show how the Bayesian approach can be used to account for model uncertainty, and in Section 6 how it resolves the difficulties with *P*-values discussed in Section 2. In Section 7 I discuss modeling strategies, and in the Appendix I describe some valuable software.

2. PRACTICAL DIFFICULTIES WITH *P*-VALUES

2.1. *P*-values

The standard statistical approach to hypothesis testing assumes that only two hypotheses, H_0 and H_1, are envisaged, and that one of these, the null hypothesis H_0, is nested within the other one. The alternative hypothesis H_1 is represented by a probability model with

d_1 unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d_1})$. H_0 can be represented by the same probability model as H_1 but with v constraints imposed on $\boldsymbol{\theta}$, $g_i(\boldsymbol{\theta}) = 0$ $(i = 1, \ldots, v)$. H_0 can represent not only exclusion restrictions such as $\theta_1 = 0$ but also linear restrictions on the parameters of H_1, such as $\theta_1 - \theta_2 = 0$ or nonlinear restrictions such as $\theta_1^2 + \theta_2^2 = 1$ (restrictions such as the latter arise in association models for contingency tables).

A test statistic T is selected and calculated from the data at hand, D; its observed value is denoted by $t(D)$. The null hypothesis H_0 is rejected in favor of the alternative hypothesis H_1 if $t(D)$ is more extreme than would be expected if H_0 were true. This is implemented by choosing a significance level α (conventionally taken to be .05 or .01), and rejecting H_0 if the probability of T being greater than or equal to $t(D)$ is small (i.e., less than α), given that H_0 is true. More formally, H_0 is rejected if

$$P = \Pr[T \geq t(D)|H_0] < \alpha, \tag{1}$$

in which case H_1 is adopted. The quantity P is called the P-value and is often reported as an indication of the strength of the evidence against H_0.

This approach is so widely applied that it is often used without its basis being critically discussed. There are, however, several features worth noting. A first point is that the significance level α has to be determined. It has become conventional to use $\alpha = .05$ or .01, based on Sir Ronald Fisher's experience with relatively small agricultural experiments (on the order of 30 to 200 plots). Subsequent advice has emphasized the need to take into account the power of the test against H_1 when setting α, and to balance power and significance in some appropriate way. However, a precise way of doing this is lacking, and this advice seems to boil down to a vague suggestion that α be lower for large sample sizes, a suggestion that is mostly ignored in practice. We will see that for the sample sizes often found in sociology, values of α much lower than the conventional ones can be appropriate.

A second point to note is that the whole standard hypothesis-testing framework rests on the basic assumption that only two models are ever entertained. This is far from being the case in most sociological studies, which are often not experimental, and where a wide range of possible control variables must be considered. In practice a selection is made among the variables, and each possible choice represents

TABLE 1
Social Mobility Tables for 16 Countries, Father's Occupation by Son's Occupation. The categories are white-collar, blue-collar, and farm.

Australia			Belgium			France			Hungary		
292	170	29	497	100	12	2085	1047	74	479	190	14
290	608	37	300	434	7	936	2367	57	1029	2615	347
81	171	175	102	101	129	592	1255	1587	516	3110	3751
Italy			**Japan**			**Philippines**			**Spain**		
233	75	10	465	122	21	239	110	76	7622	2124	379
104	291	23	159	258	20	91	292	111	3495	9072	597
71	212	320	285	307	333	317	527	3098	4597	8173	14833
United States			**West Germany**			**West Malaysia**			**Yugoslavia**		
1650	641	34	3634	850	270	406	235	144	61	24	7
1618	2692	70	1021	1694	306	176	369	183	37	92	13
694	1648	644	1068	1310	1927	315	578	2311	77	148	223
Denmark			**Finland**			**Norway**			**Sweden**		
79	34	2	39	29	2	90	29	5	89	30	0
55	119	8	24	115	10	72	89	11	81	142	3
25	48	84	40	66	79	41	47	47	27	48	29

Source: Grusky and Hauser (1983).

a different model; with p possible variables, the number of candidate models may reach 2^p, which can be huge (e.g., when $p = 15$, $w^p = 32{,}768$). We will see in Section 2.3 that failing to take into account the model selection process can yield very misleading results.

In the following sections I will outline some practical difficulties with P-value-based tests in sociological applications and give examples. I will return to the examples later in Section 6, after outlining the Bayesian approach to the problem.

2.2. *Large Samples*

Table 1 contains a three-way $3 \times 3 \times 16$ contingency table showing 3×3 social mobility tables for 16 countries, from Grusky and Hauser (1984)[1]. The total sample size ($n = 113{,}556$) is very large.

[1]Strangely enough, although these data have been much analyzed, they have never been published in the open literature. They are provided here to

TABLE 2
Fit of Models to Cross-National Social Mobility Data (n = 113,556)

Model	In G&H	Deviance	d.f.	BIC
1 Independence	Table 5, model 1	42970	64	42227
2 Lipset-Zetterberg	Text, p. 22	18390	120	16997
3 Quasi-symmetry	Table 5, model 2	150	16	−36
4 Saturated	—	0	0	0
5 Explanatory	Table 5, model 4	490	46	−43

Source: Grusky and Hauser (1984).

Two hypotheses were of central interest in this study: the hypothesis that mobility flows are the *same* in all industrialized countries (Lipset and Zetterberg 1959) and the hypothesis that the *patterns* of mobility (but not the actual amounts) are the same. This is the so-called FJH hypothesis (Featherman, Jones, and Hauser 1975), and the postulated common pattern is that of quasi-symmetry. Two other hypotheses are of interest as standards of comparison: the "baseline" hypothesis of independence between father's and son's occupation, and the hypothesis that there is no common pattern of mobility across countries.

Each of these four hypotheses can be represented by a log-linear model for the full three-way table, as explained by Grusky and Hauser (1984). The deviance and degrees of freedom for each model are shown in Table 2. Models 1, 3 and 4 form a nested sequence and so a test of one of these models against the next one takes the difference between their deviances and compares it with a χ^2 distribution with degrees of freedom equal to the difference between the degrees of freedom for the two models. Model 2 is also nested within model 3.[2]

It is clear that models 1 and 2 are unsatisfactory and should be rejected in favor of model 3.[3] By the standard test, model 3 should also be rejected, in favor of model 4, given the deviance difference of 150 on 16 degrees of freedom, corresponding to a P-value of about

facilitate reanalyses. They were first compiled by Hazelrigg and Garnier (1976), and have recently been reanalyzed by Xie (1992).

[2]The fifth model in Table 2 will be discussed below in Section 7.

[3]Strictly speaking, a test of the Lipset-Zetterberg hypothesis should involve only the nine industrialized countries in the sample, but imposing this restriction does not change the results.

10^{-120}. Grusky and Hauser (1984) nevertheless adopted model 3 because it explains most (99.7 percent) of the deviance under the baseline model of independence, it fits well in the sense that the differences between observed and expected counts are a small proportion of the total, and it makes good theoretical sense. This seems sensible, and yet is in dramatic conflict with the *P*-value-based test.

This type of conflict often arises in large samples, and hence is frequent in sociology with its survey datasets comprising thousands of cases. The main response to it has been to claim that there is a distinction between "statistical" and "substantive" significance, with differences that are statistically significant not necessarily being substantively important. I do not find this distinction to be a satisfactory panacea and believe that in most cases where the conflict has arisen, including the Grusky-Hauser study, it is due to the miscalibration of statistical significance using *P*-values, rather than to any real conflict between statistical and substantive significance. When statistical significance is properly calibrated, I have found that such a conflict rarely arises.

2.3. *Many Candidate Independent Variables*

Most sociological studies are observational and aim to infer causal relationships between a dependent variable and independent variables of interest. To minimize the possibility of observed associations being due to other variables and hence spurious, other independent variables that could induce spurious associations if they were left out are also included in the regression-type models that are used. I will call these "control variables."

But which control variables should be included? Clearly this choice should be guided by theory as far as possible. However, the theory can be somewhat weak and often produces only a rather long "laundry list" of possible control variables suggested by various theoretical arguments. This is especially the case when the study of a social phenomenon is in its early stages and the theory is still weak. Later, when an area of study has matured, the theory tends to be stronger and knowledge of which to control for tends to be firmer, based on the accumulated research of a community of investigators.

Typically, some choice is made and results with one or more subsets of the laundry list are presented. One would like to make the

choice on theoretical grounds, but there is usually little basis for this, as the theory or theories have already been used to establish the initial laundry list and often do not provide a basis for excluding variables from it. It is well known that including a control variable will not affect the estimation of the coefficient of the main independent variable of interest if the control variable is statistically independent of it or of the dependent variable. It would be nice to be able to use this fact to eliminate unnecessary control variables, but such independence usually is not known *a priori* and has to be assessed from the data.

We therefore have to fall back on statistical methods for choosing the control variables. Various methods are in common use. One is to always include the full laundry list. When this is long, however, and includes many variables that have little or no effect, the precision of estimates of parameters of interest can be hurt (e.g., Bishop, Fienberg, and Holland 1975, pp. 310–15); see Section 2.4 for an example.

Another common approach is to first fit the full model, screen the t-statistics for the parameters, remove the variables for which these are small, and then reestimate the resulting, reduced, model. I will call this the "screening" method. A further method (included in many statistical software packages) is stepwise variable regression, in which variables are added one at a time starting from the null model (forward selection), eliminated one at a time starting from the full model (backward elimination), or a mixture of the two, such as Efroymson's stepwise regression algorithm. Other methods include minimizing Mallows' C_p and maximizing the adjusted R^2; see Miller (1990) for an account of these and other variable selection methods in regression.

What these methods have in common is that they select one model out of the many possibilities, and then proceed as if that were the only model that had ever been considered. This can yield very misleading results, as pointed out by Freedman (1983), Freedman, Navidi and Peters (1988), Fenech and Westfall (1988), and Miller (1984, 1990). The reason is that by choosing among a large number of models one increases the probability of finding "significant" variables by chance alone. The sampling properties of these model selection methods (as distinct from those of the individual tests that make them up) are unknown in general, and there is little

theoretical rationale for preferring one of the methods to the others, although they often give different answers to the questions of interest; see Section 2.4.

This is clearly illustrated by a simple simulation experiment of Freedman (1983), which is similar in several respects to typical sociological studies. In his words:

> A matrix was created with 100 rows (data points) and 51 columns (variables). All the entries in this matrix were independent observations drawn from the standard normal distribution. The fifty-first column was taken as the dependent variable Y in a regression equation; the first 50 columns were taken as the independent variables $X_1, \ldots, X_{50}$. By construction, then, Y was independent of the X's. Ideally, R^2 should have been insignificant, by the standard F test. Likewise, the regression coefficients should have been insignificant, by the standard t test.

I replicated his experiment and obtained results similar to his. The data were analyzed in two ways, representing perhaps the two most common approaches to variable selection in sociology. The first way consisted of two passes. In the first pass, Y was regressed on all 50 of the X's, with the following results:

- $R^2 = 0.60$, $P = 0.09$;
- 21 coefficients out of the 50 were significant at the .25 level (i.e., $|t| > 1.15$);
- 7 coefficients out of the 50 were significant at the .05 level (i.e., $|t| > 1.99$).

Only the 21 variables whose coefficients were significant at the .25 level were included in the second pass. The results were as follows:

- $R^2 = 0.50$; $P = 0.00001$;
- 20 coefficients out of the 21 were significant at the .25 level;
- 14 coefficients out of the 21 were significant at the .05 level;
- 6 coefficients out of the 21 were significant at the .01 level.

TABLE 3
Stepwise Regression Results for Simulated Noise

Variable	Coefficient	t	P
Intercept	0.01	0.05	.956
X_8	0.30**	2.80	.006
X_{16}	−0.23*	−2 .00	.049
X_{36}	−0.23*	−2 .16	.034
X_{42}	0.34**	2.84	.006

*$P < .05$
**$P < .01$

In addition, a battery of diagnostic displays and tests (e.g., Weisberg 1985) showed no evidence of model inadequacy such as outliers, nonlinearity, heteroscedasticity or autocorrelation in the residuals.

In the words of Freedman (1983), "the results from the second pass are misleading indeed, for they appear to demonstrate a definite relationship between Y and the X's, that is, between noise and noise." Nevertheless, this sort of procedure is often followed in sociology (and laundry lists of 50 variables are not atypical), and many a social researcher would feel confident about presenting such findings.

Stepwise regression does not help. Table 3 shows the results: a four-variable model with $R^2 = 0.18$ and $P = 10^{-6}$, and coefficients that are all significant at the .05 level (with two also significant at the .01 level). The minimum C_p and adjusted R^2 methods also lead to models with too many predictors and highly significant F statistics.

2.4. *Model Uncertainty*

When many models are initially considered, it often happens that several of them fit the data almost equally well, or that different models are arrived at by different model selection methods. It can then happen that different models, all of them defensible, lead to different answers to the main questions of interest.

The analyst then has three main options. The first is to pick one model and adopt the conclusions that flow from it rather than from the other defensible models; this is somewhat arbitrary. The second option is to present the analyses based on all the plausible

models without choosing between them; while not fully satisfactory, this seems better than the first option. The third possibility, which I will develop in later sections, is to take account explicitly of model uncertainty when drawing conclusions.

To show how the problem can arise, consider the criminological study by the economist Isaac Ehrlich (1973), which was one of the earliest systematic efforts to determine whether greater punishments reduce overall crime rates. Up to the 1960s, criminal behavior was traditionally viewed as deviant and linked to the offender's presumed exceptional psychological, social, or family circumstances. Becker (1968) and Stigler (1970) argued, on the contrary, that the decision to engage in criminal activity is a rational choice determined by its costs and benefits relative to other (legitimate) opportunities. Ehrlich (1973) developed this argument theoretically, specified it mathematically, and tested it empirically using aggregate data from 47 U.S. states in 1960. Errors in Ehrlich's empirical analysis were corrected by Vandaele (1978), who gave the corrected data, which we use here.[4]

Ehrlich's theory goes as follows. The costs of crime are related to the probability of imprisonment and the average time served in prison, which in turn are influenced by police expenditures, which may themselves have an independent deterrent effect. The benefits of crime are related to both the aggregate wealth and income inequality in the surrounding community. The expected net payoff from alternative legitimate activities is related to educational level and the availability of employment, the latter being measured by the unemployment and labor force participation rates. This payoff was expected to be lower (in 1960) for nonwhites and for young males than for others, so that states with high proportions of these were expected also to have higher crime rates. Vandaele (1978) also included an indicator variable for southern states, the sex ratio, and state population as control variables.

We thus have 15 candidate predictors of crime rate (Table 4), and so potentially $2^{15} = 32{,}768$ different models. As in the original analyses, all analyses were done in terms of the natural logarithms of

[4]Ehrlich's study has been much criticized (e.g., Brier and Fienberg 1980) and here I use it purely as an illustrative example. For economy of expression, I will use causal language and speak of "effects," even though the validity of this language for these data is dubious.

TABLE 4
Variables in Crime Data

	Variable
1	Percent of males 14–24
2	Indicator variable for southern state
3	Mean years of schooling
4	Police expenditure in 1960
5	Police expenditure in 1959
6	Labor force participation rate
7	Number of males per 1000 females
8	State population
9	Number of nonwhites per 1000 people
10	Unemployment rate of urban males 14–24
11	Unemployment rate of urban males 35–39
12	GDP
13	Income inequality
14	Probability of imprisonment
15	Average time served in state prisons

the variables. Standard diagnostic checking did not reveal any striking violations of the assumptions underlying normal linear regression.

Interest focuses on the significance and size of the coefficients for variables 14 and 15, respectively the probability of imprisonment and the average time served in state prisons. Ehrlich (1973) did not use statistical model selection methods but instead analyzed two regression models chosen in advance on theoretical grounds.

Table 5 shows results from six models selected using methods discussed so far. The statistically chosen models 2, 3, and 4 all give high and similar values of R^2 and share many of the same variables, while Ehrlich's theoretically chosen models 5 and 6 fit less well. There are striking differences, indeed conflicts, between the results from different models. Even the statistically chosen models, despite their superficial similarity, lead to conflicting conclusions about the main questions of interest.

Consider first the effect of X_{14}, the probability of imprisonment, on the crime rate. All analyses and models concur in saying that this does have an effect, so interest focuses on estimating its size. To aid interpretation, recall that all variables have been logged, so that $\beta_{14} = -.30$ means roughly that a 10 percent increase in the

TABLE 5
Models Selected for the Crime Data

	Method	Variables	R^2 (%)	# vars.	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$	P_{15}
1	Full model	All	87	15	−.30	−.27	.133
2	Stepwise regression	1,3,4,9,11,13,14	83	7	−.19	—	—
3	Mallows' C_p	1,3,4,9,11,12,13,14,15	85	9	− .30	−.30	.050
4	Adjusted R^2	1,3,4,7,8,9,11,12,13,14,15	86	11	−.30	−.25	.129
5	Ehrlich model 1	9,12,13,14,15	66	5	−.45	−.55	.009
6	Ehrlich model 2	1,6,9,10,12,13,14,15	70	8	−.43	−.53	.011

Note: P_{15} is the P-value from a two-sided t-test for testing $\beta_{15} = 0$.

probability of imprisonment produces a 3 percent reduction in the crime rate, all else being equal. The estimates of β_{14} fluctuate wildly between models. The stepwise regression model gives an estimate that is about one-third lower in absolute value than the full model, a difference that may be large enough to be of policy importance; this difference is equal to about 1.7 standard errors. The Ehrlich models give estimates that are about one-half higher than the full model, and more than twice as big as those from stepwise regression (in absolute value). There is clearly considerable model uncertainty about this parameter.

Another point of interest, not shown in Table 5, is that the standard error of $\hat{\beta}_{14}$ (and also of the other coefficients) is smaller for the more parsimonious models. For the full model, it is .098, while for the stepwise regression model it is .066. Thus it could be argued that retaining the additional nonsignificant variables in the full model reduces the efficiency of estimation of β_{14} by a factor of $(.066/.098)^2 = .45$, and so is equivalent to throwing away more than half the data.

Now let us turn to β_{15}, the effect of the average time served in state prisons. Whether this is significant at all is not clear, and t-tests based on different models lead to different conclusions. In the full model it has a nonsignificant P-value of .133, while stepwise regression leads to a model that does not include the variable at all. On the other hand, Mallows's C_p leads to a model in which it is just significant at the .05 level, while with adjusted R^2 it is again not significant. In Ehrlich's models, by contrast, it is highly significant.

Together these results paint a confused picture about β_{15}, and there seem to be no frequentist results to help sort it out. I will argue that the confusion can be resolved by taking account explicitly of the model uncertainty.

2.5. *Nonnested Hypotheses, and Evidence* for *the Null Hypothesis*

Often, in sociology, competing hypotheses represent quite different views of the phenomenon being studied and cannot easily be neatly represented by nested statistical models. For instance, in the crime example of the preceding section, one hypothesis might be that criminal behavior is deviant and explainable by the criminal's own characteristics, while a competing hypothesis would be that it is a rational choice. Adjudicating between such hypotheses often involves com-

paring nonnested models, and so the standard theory of Section 2.1 breaks down.

One way around this has been proposed by Cox (1961, 1962); it has been applied to sociological problems by Weakliem (1992) and Halaby and Weakliem (1993). Cox's approach, which has spawned a large literature, tends to be cumbersome to implement and requires the often arbitrary designation of one of the two nonnested models as the null hypothesis. One way around this arbitrariness is to carry out two tests rather than one, with each model in turn as the null hypothesis. However, there is no guarantee of getting the standard kind of result of a test, namely rejection of one model and nonrejection of the other. Both models may fail to be rejected, in which case it is not clear how to make inferences about quantities of interest, especially if the two models lead to different conclusions. Both models may be rejected (as often happens with large samples), in which case the tests do not provide a comparison between the two models.

Another difficulty is that standard significance tests allow one either to reject the null hypothesis or to fail to reject it, but they do not provide any measure of evidence *for* the null hypothesis. Sometimes, however, sociological theories specify that something is the *same* across different groups, and thus the null hypothesis is the hypothesis of interest. One example is the Lipset-Zetterberg hypothesis referred to earlier in Section 2.2, that social mobility flows are the same in all industrialized countries. Another is the hypothesis that all sections of U.S. society now obey a two-child norm, according to which most couples have two children and there is very little variation between socioeconomic groups in average completed family size (among those who have any children) (Lye and Greek, 1994).

A standard test allows us to say only that the data have failed to reject our null hypothesis of interest but gives no indication of whether the data support it or not. A test can fail to reject a null hypothesis either because there is not enough data, or because the data do support it, but it does not allow us to distinguish between these two different situations.

Difficulties with P-values and the associated significance tests have been much discussed in the scientific literature. The reader edited by Morrison and Henkel (1970) compiled about 30 important pre-1970 articles, the majority of them by sociologists; they are still

worth reading. They referred a great deal to the problems with large samples, but talked very little about the other difficulties discussed here; they did not suggest alternatives that would seem fully satisfactory nowadays. Leamer (1978) was the first to discuss in depth the difficulties with empirical model-building using significance tests. Recent social science references include Johnstone (1990*a*, *b*).

3. BAYESIAN HYPOTHESIS TESTING

In this section, I first briefly review Bayesian statistical parameter estimation, and then introduce Bayes factors, which form the basis for Bayesian hypothesis testing.

3.1. *Bayesian Estimation*

Bayesian estimation expresses all uncertainty, including uncertainty about the unknown parameters of a model, in terms of probability, and it views unknown parameters as random variables. Thus all results in Bayesian statistics follow directly from elementary probability theory, notably the definition of conditional probability, Bayes' theorem, and the law of total probability.

We start with a probability model for the data D, which is specified by a vector of d unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. Before any data are observed, our beliefs and uncertainty about $\boldsymbol{\theta}$ are represented by a prior probability density $p(\boldsymbol{\theta})$. The probability model is specified by the likelihood $p(D|\boldsymbol{\theta})$, which is the probability of observing the data D given that $\boldsymbol{\theta}$ is the true parameter.

Having observed the data D, we update our beliefs about $\boldsymbol{\theta}$ using Bayes' theorem to obtain the posterior distribution of $\boldsymbol{\theta}$ given the data D, namely

$$p(\boldsymbol{\theta}|D) = p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(D), \tag{2}$$

where $p(D) = \int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, by the law of total probability. For estimation purposes we need to know $p(\boldsymbol{\theta}|D)$ only up to a constant of proportionality, and since $p(D)$ does not involve $\boldsymbol{\theta}$ it can be omitted from equation (2), which is then written

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3}$$

Thus the posterior distribution is proportional to the likelihood times the prior.

The posterior distribution, $p(\boldsymbol{\theta}|D)$, contains all the information needed to make inference about $\boldsymbol{\theta}$. The only question is how best to summarize and communicate that information. Often interest focuses on the individual parameters (i.e., the components of $\boldsymbol{\theta}$). The posterior distribution of a component of $\boldsymbol{\theta}$, say θ_1, follows from the law of total probability by *integrating* out the other components, so that

$$p(\theta_1|D) = \int p(\boldsymbol{\theta}|D)d\theta_2 d\theta_3 \ldots d\theta_d. \tag{4}$$

The univariate distribution (4) contains all the information needed to make inferences about θ_1. It can be summarized in various ways. In my experience, the mosts useful summaries are the posterior mode—i.e., the value of θ_1 that maximizes $p(\theta_1|D)$ and so is the most likely value given the data—and the .025 and .975 quantiles, which define a 95 percent Bayesian confidence interval. The posterior standard deviation is also useful, as a Bayesian analogue of the standard error. The posterior mean is also often used and is usually close to the posterior mode.

Bayesian inference has been controversial because it uses the prior distribution, $p(\boldsymbol{\theta})$, which is subjectively determined by the user. However, in large samples this has very little influence: Its contribution to the posterior mean and variance is on the order of $(1/n)$-th of the total, where n is the sample size.

In large samples, the posterior mode is very close to the maximum likelihood estimator (MLE), and Bayesian confidence intervals are very similar to standard non-Bayesian confidence intervals. Asymptotically, in regular models,[5] the posterior distribution is multivariate normal with mean at the MLE and variance matrix equal to the inverse (observed or, less accurately, expected) Fisher information matrix. Thus, for *estimation* in *regular models* with *large samples,* Bayesian and maximum likelihood methods give answers that are essentially the same. The answers can be different, however, for testing and model selection, for estimation in nonregular models, and with very small samples.

[5]A regular statistical model is one in which the MLE is asymptotically normal with mean at the true value and variance matrix equal to the inverse expected Fisher information matrix. A simple example of a nonregular model is that in which the data are independent and uniformly distributed between 0 and θ, and θ is unknown. Then the MLE of θ is equal to the largest observation and does not have the usual asymptotic distribution (Kotz and Johnson 1985, p. 346).

Edwards, Lindman and Savage (1963) gave what remains an excellent and delightfully written introduction to Bayesian statistics for a social science audience, while Press (1989) and Lee (1989) are accessible accounts in book form. For a more advanced and theoretical treatment, but one that is still practically motivated, see Bernardo and Smith (1994).

3.2. *Bayes Factors*

Suppose now that we want to use the data D to compare two competing hypotheses, which are represented by the statistical models M_1 and M_2, with parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. They may be nested, but need not be. Then, by Bayes' theorem, the posterior probability that M_1 is the correct model (given that either M_1 or M_2 is) is

$$p(M_1|D) = \frac{p(D|M_1)p(M_1)}{p(D\,|M_1)p(M_1) + p(D|M_2)p(M_2)}, \tag{5}$$

where $p(D|M_k)$ is the (marginal) probability of the data given M_k (see below), and $p(M_k)$ is the prior probability of model M_k ($k = 1,2$). A similar expression holds for $p(M_2|D)$ and, by construction, $p(M_1|D) + p(M_2|D) = 1$.

In equation (5), $p(D|M_1)$ is obtained by *integrating* (not maximizing) over $\boldsymbol{\theta}_1$, i.e.,

$$\begin{aligned} p(D|M_1) &= \int p(D|\boldsymbol{\theta}_1, M_1)p(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_1 \\ &= \int (\text{likelihood} \times \text{prior})d\boldsymbol{\theta}_1, \end{aligned} \tag{6}$$

where $p(D|\boldsymbol{\theta}_1, M_1)$ is the likelihood of $\boldsymbol{\theta}_1$ under model M_1. I will call this quantity, $p(D|M_1)$, the *integrated likelihood* for model M_1; it has also been called the marginal likelihood, the marginal probability of the data, and the predictive probability of the data.

The extent to which the data support M_2 over M_1 is measured by the *posterior odds* for M_2 against M_1—that is, the ratio of their posterior probabilities. By equation (5), this is

$$\frac{p(M_2|D)}{p(M_1|D)} = \left[\frac{p(D|M_2)}{p(D|M_1)}\right]\left[\frac{p(M_2)}{p(M_1)}\right]. \tag{7}$$

The first factor on the right-hand side of equation (7) is the ratio of the integrated likelihoods of the two models and is called the *Bayes factor* for M_2 against M_1, denoted by B_{21}. The second factor on the right-hand side of (7) is the prior odds, and this will often be equal to 1, representing the absence of a prior preference for either model—that is, $p(M_1) = p(M_2) = \frac{1}{2}$. Thus equation (7) can be written

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds}. \tag{8}$$

It follows that the Bayes factor is equal to the posterior odds when the prior odds are equal to 1.

When $B_{21} > 1$, the data favor M_2 over M_1, and when $B_{21} < 1$ the data favor M_1. The use of Bayes factors to compare scientific theories was first proposed by Jeffreys (1935), and in 1961 he proposed the following rules of thumb for interpreting B_{21} (Jeffreys 1961, Appendix B): When $1 \leq B_{21} \leq 3$, there is evidence for M_2, but it is "not worth more than a bare mention," when $3 \leq B_{21} \leq 10$ the evidence is positive, when $10 \leq B_{21} \leq 100$ it is strong, and when $B_{21} > 100$ it is decisive. Probability itself is a meaningful scale and so these categories are not a calibration of the Bayes factor but rather a rough descriptive statement about standards of evidence in scientific investigation. I will return to the issue of interpretation in Section 4.3 and suggest a slightly different scale for use in social research.

Evaluating the Bayes factor involves calculating the integrated likelihood (6), which can be a high-dimensional and intractable integral. Various analytic and numerical approximations have been proposed, and in Section 4 I will discuss the BIC approximation, which is both simple and accurate. The Bayes factor depends on the prior and, in principle, this should be carefully specified and sensitivity to it should be carefully assessed. However, as we will see in Section 4.1, the BIC approximation corresponds rather closely to a particular choice of prior that seems reasonable for many practical purposes.

These and other aspects of Bayes factors are reviewed in detail by Kass and Raftery (1995), who give many references. One point they make is that the logarithm of the integrated likelihood may also be viewed as a predictive score for the model (Kass and Raftery, 1995, Section 3.2). This is of interest because it leads to an interpretation of the Bayes factor that does not depend on viewing one of the models as

"true." In this view, the Bayes factor is designed to choose the model that will, on average, give better out-of-sample predictions.

4. THE BIC APPROXIMATION

In this section, I will introduce the BIC (*Bayesian Information Criterion*) approximation to the Bayes factor by deriving it heuristically, giving explicit expressions for it in various model classes, and finally discussing its interpretation and its relation to *P*-values.

4.1. *Derivation*

The key quantity underlying the Bayes factor is the integrated likelihood for a model, given by equation (6). I will first derive a simple approximation to this quantity, and then show how it leads to approximate Bayes factors and to the BIC criterion for assessing models. This subsection is fairly technical. The key result is equation (20) and, if you are not interested in the derivation of BIC, you can now skip to that point and still be able to follow the rest of the chapter.

For the moment I will concentrate on approximating the integrated likelihood for a single model, and for simplicity I will simplify notation by not mentioning the model, so that equation (6) will be rewritten

$$p(D) = \int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{9}$$

For ease of exposition, I will consider the case where the data D consist of n independent and identically distributed observations, $y_1, \ldots, y_n$, each of which may be a vector. The results apply much more widely than this, however, and in essence are valid for any regular statistical model. This includes many time-series models for data that are not independent, and also models for data that are not identically distributed. For example, it includes most common models for event-history data.

The derivation proceeds by considering a Taylor series expansion of $g(\boldsymbol{\theta}) = \log\{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})\}$ about $\tilde{\boldsymbol{\theta}}$, the value of $\boldsymbol{\theta}$ that maximizes $g(\boldsymbol{\theta})$, i.e. the posterior mode. The expansion is

$$\begin{aligned} g(\boldsymbol{\theta}) = g(\tilde{\boldsymbol{\theta}}) &+ (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g'(\tilde{\boldsymbol{\theta}}) + \tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\ &+ o(\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|^2), \end{aligned} \tag{10}$$

where the superscript ^{T} denotes matrix transpose, $g'(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_d}\right)^T$ is the vector of first partial derivatives of $g(\boldsymbol{\theta})$, and $g''(\boldsymbol{\theta})$ is the Hessian matrix of second partial derivatives of $g(\boldsymbol{\theta})$ whose (i, j) element is $\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}$. Now $g'(\tilde{\boldsymbol{\theta}}) = 0$ because $g(\boldsymbol{\theta})$ reaches a maximum at $\tilde{\boldsymbol{\theta}}$ and so its first derivative is equal to zero at that point. Thus

$$g(\boldsymbol{\theta}) \approx g(\tilde{\boldsymbol{\theta}}) + \tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}). \tag{11}$$

The approximation in equation (11) is not sure to be good unless $\boldsymbol{\theta}$ is close to $\tilde{\boldsymbol{\theta}}$. However, when n is large, the likelihood $p(D|\boldsymbol{\theta})$ is concentrated about its maximum and declines fast as one moves away from $\tilde{\boldsymbol{\theta}}$, so that only values of $\boldsymbol{\theta}$ close to $\tilde{\boldsymbol{\theta}}$ will contribute much to the integral (9) defining $p(D)$. For a formalization of this argument see Tierney and Kadane (1986).

It follows that

$$\begin{aligned} p(D) &= \int \exp[g(\boldsymbol{\theta})]d\boldsymbol{\theta} \\ &\approx \exp[g(\tilde{\boldsymbol{\theta}})] \int \exp\left[\tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\right]d\boldsymbol{\theta}, \end{aligned} \tag{12}$$

by equation (11). Recognizing the integrand in equation (12) as proportional to a multivariate normal density gives

$$p(D) \approx \exp[g(\tilde{\boldsymbol{\theta}})](2\pi)^{d/2}|A|^{-1/2}, \tag{13}$$

where d is the number of parameters in the model and $A = -g''(\tilde{\boldsymbol{\theta}})$. The use of equation (13) is called the *Laplace method for integrals*. The error in equation (13) is $O(n^{-1})$ (Tierney and Kadane, 1986), and so

$$\log p(D) = \log p(D|\tilde{\boldsymbol{\theta}}) + \log p(\tilde{\boldsymbol{\theta}}) + (d/2)\log(2\pi) - \tfrac{1}{2}\log|A| + O(n^{-1}), \tag{14}$$

where $O(n^{-1})$ represents any quantity such that $nO(n^{-1}) \to$ a constant as $n \to \infty$.

Now in large samples, $\tilde{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ is the MLE, and $A \approx n\mathbf{i}$, where $\mathbf{i}$ is the expected Fisher information matrix for one observation. This is a $(d \times d)$ matrix whose (i, j) element is $-E\left[\frac{\partial^2 \log p(y_1|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]$, the expectation being taken over values of y_1, with $\boldsymbol{\theta}$ held fixed. Thus $|A| \approx n^d|\mathbf{i}|$. These two approximations introduce an $O(n^{-1/2})$ error into equation (14), which becomes

$$\log p(D) = \log p(D|\hat{\boldsymbol{\theta}}) + \log p(\hat{\boldsymbol{\theta}}) + (d/2)\log(2\pi) - (d/2)\log n - \tfrac{1}{2}\log|\mathbf{i}| + O(n^{-1/2}). \tag{15}$$

Now the first term on the right-hand side of equation (15) is of order $O(n)$, the fourth term is of order $O(\log n)$, while the other four terms are of order $O(1)$ or less. Removing the terms of order $O(1)$ or less thus gives

$$\log p(D) = \log p(D|\hat{\boldsymbol{\theta}}) - (d/2)\log n + O(1). \tag{16}$$

Equation (16) says that the log-integrated likelihood, $\log p(D)$, is equal to the maximized log-likelihood, $\log p(D|\hat{\boldsymbol{\theta}})$, minus a correction term.

Equation (16) is the approximation on which BIC is based, and its $O(1)$ error means that, in general, the error in it does not vanish even with an infinite amount of data. This is not as bad as it sounds, however, because the other terms on the right-hand side of (16) tend to infinity as n does, and so will eventually dominate. Thus the error in (16) will tend toward zero as a *proportion* of $\log p(D)$, ensuring that the error will not affect the conclusion reached, given enough data. Nevertheless, the $O(1)$ error does suggest the approximation to be somewhat crude.

Empirical experience has found (16) to be more accurate in practice than the $O(1)$ error term would suggest (e.g., Raftery 1993*b*). In fact, the error is of a much smaller order of magnitude for a particular, reasonable, choice of prior distribution. Suppose that the prior $p(\boldsymbol{\theta})$ is multivariate normal with mean $\hat{\boldsymbol{\theta}}$ and variance matrix $\mathbf{i}^{-1}$. Thus, roughly speaking, the prior distribution contains the same amount of information as would, on average, a single observation. This seems to be a reasonable representation of the common situation where there is a little, but not much, prior information. Then

$$\log p(\hat{\boldsymbol{\theta}}) = -(d/2)\log(2\pi) + \tfrac{1}{2}\log|\mathbf{i}|, \tag{17}$$

and substituting (17) into (15) gives

$$\log p(D) = \log p(D|\hat{\boldsymbol{\theta}}) - (d/2)\log n + O(n^{-1/2}). \tag{18}$$

Thus for the particular prior mentioned, the error in the approximation (16) is $O(n^{-1/2})$ rather than $O(1)$, which is much smaller for moderate to large sample sizes, and which does tend to zero as n tends to infinity.

The approximation (18) can be used to approximate the Bayes factor $B_{21} = p(D|M_2)/p(D|M_1)$. This is most conveniently written on the scale of twice the logarithm, as follows:

$$2 \log B_{21} = 2 (\log p(D|\hat{\boldsymbol{\theta}}_2, M_2) - \log p(D|\hat{\boldsymbol{\theta}}_1, M_1)) - (d_2 - d_1) \log n + O(n^{-1/2}). \quad (19)$$

If M_1 is nested within M_2, equation (19) can be rewritten

$$2 \log B_{21} \approx \chi^2_{21} - \mathrm{df}_{21} \log n, \quad (20)$$

where χ^2_{21} is the standard likelihood ratio test (LRT) statistic for testing M_1 against M_2, and $\mathrm{df}_{21} = d_2 - d_1$ is the number of degrees of freedom associated with the test.

The Laplace method for integrals was introduced into statistics by Tierney and Kadane (1986) and seems first to have been used for Bayes factors by Raftery (1988). Equation (15) goes back to Jeffreys (1961), while equation (16) is due to Schwarz (1978) and equation (18) was pointed out by Kass and Wasserman (1992). For other references, see Kass and Raftery (1995).

4.2. *BIC for Specific Models*

4.2.1. *General Form*

When several models are being considered, it is useful to compare each of them in turn with a baseline model, usually either a null model (M_0) with no independent variables, or a saturated model (M_S) in which each data point is fit exactly.

When the baseline model is a saturated model, M_S, the LRT statistic in equation (20) is often called the *deviance*. The value of BIC for model M_k, denoted by BIC_k, is the approximation to $2 \log B_{Sk}$ given by (20), where B_{Sk} is the Bayes factor for model M_S against model M_k. This is

$$\mathrm{BIC}_k = L^2_k - df_k \log n, \quad (21)$$

where $L^2_k = \chi^2_{Sk}$ is the deviance for model M_k and df_k is the corresponding number of degrees of freedom. Then BIC_S, the BIC value for the saturated model, is zero, and the saturated model is preferred to M_k if $\mathrm{BIC}_k > 0$, in which case M_k can be considered not to fit the data well. When $\mathrm{BIC}_k < 0$, M_k is preferred to the saturated model, and the smaller that BIC_k is (i.e., the more negative), the better the fit of M_k.

When comparing two models, M_j and M_k, we note that

$$
\begin{aligned}
B_{jk} &= p(D|M_j)/p(D|M_k) \\
&= \left[\frac{p(D|M_S)}{p(D|M_k)}\right] \Big/ \left[\frac{p(D|M_S)}{p(D|M_j)}\right] \\
&= B_{Sk}/B_{Sj}, \text{ and so} \\
2\log B_{jk} &= 2\log B_{Sk} - 2\log B_{Sj} \\
&\approx \mathrm{BIC}_k - \mathrm{BIC}_j. \qquad (22)
\end{aligned}
$$

Thus two models can be compared by taking the difference of their BIC values, with the model having the smaller (i.e., the more negative) BIC value being preferred. I will discuss the interpretation of the size of the difference in Section 4.3. Note that M_j and M_k do not have to be nested for equation (22) to be applicable.

When the baseline model is the null model, M_0, with no independent variables, then BIC_k is replaced by BIC'_k, the approximation (20) to $2\log B_{0k}$, where B_{0k} is the Bayes factor for the null model M_0 against the model of interest M_k. This is

$$\mathrm{BIC}'_k = -\chi^2_{k0} + p_k \log n, \qquad (23)$$

where χ^2_{k0} is the LRT statistic for testing M_0 against M_k, and p_k is the number of degrees of freedom associated with that test. In regression-type models, p_k will usually be the number of independent variables in M_k.

BIC'_0, the BIC′ value for the null model, is zero. Thus if BIC'_k is positive, the null model M_0 is preferred to M_k, indicating that M_k is overparameterized, containing parameters (and hence probably variables) for which the data provide little support. In that case, a submodel of M_k (containing some but not all of the variables in M_k) may well fit better than either M_0 or M_k. For examples of this, see Section 7. If BIC'_k is negative, then M_k is preferred to M_0, and the smaller (i.e., the more negative) BIC'_k is, the more M_k is supported by the data. For comparing two models, BIC′ differences can be used in the same way as BIC differences, and equation (22) is still valid if BIC is replaced by BIC′.

Which of BIC_k or BIC'_k should be used? For any one model, they will be numerically different, but for *comparing* any two given models, M_j and M_k, they are equivalent, in the sense that the BIC difference is the same as the BIC′ difference, i.e.

$$\mathrm{BIC}_k - \mathrm{BIC}_j = \mathrm{BIC}'_k - \mathrm{BIC}'_j. \tag{24}$$

The two measures, BIC and BIC′, differ only by a constant that is the same for all models; this constant is equal to both BIC_0 and $-\mathrm{BIC}'_S$, which are in turn equal to one another. Thus

$$\mathrm{BIC}_k - \mathrm{BIC}'_k = c \tag{25}$$

for all models M_k, where $c = \mathrm{BIC}_0 = -\mathrm{BIC}'_S$.

In practice, which of BIC or BIC′ is used will depend on whether the software that estimates the models provides the deviances or the LRT statistic against the null model. If the software yields the deviance, then BIC will be used, and if instead it reports the LRT statistic, then BIC′ will be used. If both the deviance and the LRT statistic are available, either BIC or BIC′ can be used, or both. Although equivalent for testing and model selection purposes, they do each provide some different information. BIC_k can be viewed as a measure of overall model fit,[6] while BIC'_k provides an assessment of whether M_k is explaining enough of the variation in the data to justify the number of parameters it uses.

There is one important ambiguity in equations (21) and (23)—namely, the definition of n, the "sample size." What this should be is clear in some situations but not in others. As a general rule, the definition of n should be the one that makes the approximation $|A| \approx n^d|\mathbf{i}|$ used in the derivation of (15) most accurate. More precise suggestions for specific model classes will be given in the following subsections.

4.2.2. *Linear Regression and Analysis of Variance*

For linear regression with normal errors, the most convenient form is BIC′, and it can be shown that this has the simple form

$$\mathrm{BIC}'_k = n \log(1 - R_k^2) + p_k \log n, \tag{26}$$

where R_k^2 is the value of R^2 for model M_k and p_k is the number of independent variables (not including the intercept).

Note that standard analysis of variance for designed experiments can be recast in terms of linear regression by using sets of dummy variables to represent the different factors and interactions,

[6]This is true only in models for which goodness-of-fit statistics can be used for this purpose, such as models for categorical data.

and then equation (26) can be used in that context also. In particular, simple problems like testing for a difference between two means can be solved using (26) in this way.

The sample size n will usually be just the number of cases. This will not be true, however, if responses with the same values of the independent variables have been grouped into a single case with the average response as dependent variable, and weighted regression carried out, with weights proportional to the number of individuals in the group. This often happens in the analysis of designed experiments, when individuals are grouped into "cells." The n should be the actual number of individuals rather than the number of cases or cells. When the data have been collected using a complex survey design with resulting weights, it is not yet clear what n should be, and this issue awaits further study. However, it seems reasonable that if the model is based on an assumption of simple random sampling but the sampling design is less efficient, then n should be reduced to reflect the efficiency of the sampling design relative to simple random sampling.

4.2.3. *Logistic Regression*

Some logistic regression software produces the deviance, some the LRT statistic, and some both. Thus BIC and BIC′ may both be used, depending on the software, and equations (21) and (26) apply directly. The same is true for other binary response models, such as those with the probit or complementary log-log link.

What should n be? When each individual is a separate case, then n should be simply the sample size. In logistic regression, however, responses with the same values of the independent variables are often grouped together into a single case for which the dependent variable is the number of positive responses, which has a binomial distribution. In that situation, the number of cases is not the same as the number of individuals. Then n should be the number of individuals—i.e., the sum of the binomial denominators, not the number of cases in the regression.

4.2.4. *Log-Linear Modeling*

Software that estimates log-linear models for contingency tables usually gives the deviance rather than the LRT statistic against a null model. Thus it is most natural to use BIC rather than BIC′.

What should n be? Once again, it is best to use the actual number of individuals—i.e., the sum of the cell counts, *not* the number of cells (Raftery 1986*a*).

4.2.5. *Event-History Analysis*

Most event-history analysis software reports the LRT statistic against the null model with no independent variables, and so BIC′ is the more convenient measure to use. For fully parametric event-history models, the theory of Section 4.1 provides a direct justification for the use of BIC′. However, event-history analysis is often based on the Cox proportional hazards model, and there there is a complication: It is not fully parametric because the baseline hazard rate is unspecified. The regression part *is* parametric, however, and this is a case of a semiparametric model. In spite of this, BIC′ may still be validly used for the Cox model (Raftery, Madigan, and Volinsky 1995). The number of degrees of freedom, p_k, is then just the number of independent variables.

What should n be? Should it be the number of individuals, the number of events, or the number of spells (including censored spells)? It seems best to use the number of events rather than either of the other two possibilities (Raftery, Madigan, and Volinsky 1995).

For discrete-time event-history analysis, the same choice has been made (Xie 1994), while the total number of exposure time units has also been used, for consistency with logistic regression (Raftery, Lewis, Aghajanian and Kahn 1995; Raftery, Lewis and Aghajanian 1995). The latter choice is more conservative and seems safer in the absence of a definitive result about which is more appropriate. More research is needed on this matter, and I conjecture that the less conservative choice of Xie (1994) will eventually be shown to be the more appropriate one.

4.2.6. *Structural Equation Models*

In this subsection I will use the notation of Bollen (1989, table 2.2), so that N is the number of individuals, p is the number of indicators of the independent variables, q is the number of indicators of the dependent variables, and v_k is the number of independent parameters fitted in model M_k.

Software for estimating structural equation models, such as LISREL or EQS, tends to give the deviance (i.e., the LRT statistic

against the "saturated" model in which each *covariance* is fit exactly), rather than the LRT statistic against a null model. Thus BIC rather than BIC′ is the more convenient measure and equation (21) is the one to use. There df_k is the number of covariances minus the number of independent parameters fitted—that is, $df_k = \frac{1}{2}(p + q)(p + q + 1) - v_k$.

When one is comparing two models, M_k and M_{k-1}, where M_{k-1} is nested within M_k and M_k has one more parameter than M_{k-1} then, approximately, $L^2_{k-1} - L^2_k = t^2$, where L^2_k is the deviance for model M_k and t is the t test statistic for testing the additional parameter. Thus

$$\text{BIC}_{k-1} - \text{BIC}_k \approx t^2 - \log n. \tag{27}$$

If this is positive, the larger model M_k will be preferred.

When one is comparing M_k with a bigger model, M_{k+1} within which it is nested and which has one *more* parameter than M_k, then, approximately, $L^2_k - L^2_{k+1} = W$, the Lagrange multiplier test statistic or modification index, and so

$$\text{BIC}_k - \text{BIC}_{k+1} \approx W - \log n. \tag{28}$$

Again, if this is positive, the larger model M_{k+1} will be preferred.

Equations (27) and (28) are useful for model-building with BIC in structural equation models, because most software for estimating these models returns both t statistics and modification indices. Thus by fitting a single model, one can compute approximate BIC values for it, all the models that have one parameter less than it, and all the models that have one parameter more than it. For an example of a model search that exploits this fact, see Raftery (1993*a*).

What should n be? I recommend using $n = N$, the number of individuals. Raftery (1993*a*) used $n = N(p + q)$, but the derivation of equation (19) (which was not known when Raftery [1993*a*] was written) suggests that $n = N$ would be more accurate. Note, however, that equation (16) is valid for both definitions of n.

4.3. *Interpretation*

In Section 3.2 I gave the rules of thumb of Jeffreys (1961) for interpreting Bayes factors and, hence, between-model differences in BIC or BIC′. I find a slightly modified version more appropriate. I prefer

TABLE 6
Grades of Evidence Corresponding to Values of the Bayes Factor for M_2 Against M_1, the BIC Difference and the Posterior Probability of M_2

BIC Difference	Bayes Factor	$p(M_2\|D)(\%)$	Evidence
0–2	1–3	50–75	Weak
2–6	3–20	75–95	Positive
6–10	20–150	95–99	Strong
>10	>150	>99	Very strong

to define "strong" evidence as corresponding to posterior odds of 20:1 rather than 10:1 (by analogy with the intention behind the standard .05 significance level), and to use the term "very strong" rather than "decisive" for the evidence implied by very high posterior odds. Jeffreys put the boundary for this at 100:1, corresponding to a BIC difference of 2 log 100 = 9.2, but I prefer to round this up to the slightly more conservative value of 10, corresponding to posterior odds of about 150:1. This yields the scheme shown in Table 6.

A conversion of t statistics and their associated P-values to approximate BIC differences can be made by noting that when $df_{21} = 1$ in equation (20), then, approximately in regular models, $\chi^2_{21} \approx t^2$, where t is the usual t statistic for testing the significance of the parameter of M_2 that is set equal to zero in M_1. Then (20) becomes

$$2\log B_{21} \approx t^2 - \log n \approx \mathrm{BIC}_1 - \mathrm{BIC}_2. \tag{29}$$

(Note that the middle expression in equation [29] is only an approximation to the difference of BIC values—that is, an approximation to an approximation.) It follows that t values can be roughly translated into BIC values and hence into grades of evidence such as those of Table 6. In particular, $|t| > \sqrt{\log n}$ is required for there to be even weak evidence for the additional parameter in M_2, while $|t| > \sqrt{\log n + 6}$ corresponds to strong evidence on this scale.

Table 7 shows the minimum t values required for various grades of evidence and sample sizes. The sample sizes are chosen to represent roughly the sample sizes that arise in various kinds of sociological study. The first three sample sizes are in the range of those that arise in aggregate studies and in quantitative macrosociology: very roughly, there are about 30 industrialized countries, 50 U.S. states, and 100 U.S. SMSAs in a typical study. The last

TABLE 7
Approximate Minimum t Values Corresponding to Different Grades of Evidence

Evidence	Minimum BIC Difference	n					
		30	50	100	1,000	10,000	100,000
Weak	0	1.84	1.98	2.15	2.63	3.03	3.39
Positive	2	2.32	2.43	2.57	2.98	3.35	3.68
Strong	6	3.07	3.15	3.26	3.59	3.90	4.18
Very strong	10	3.66	3.73	3.82	4.11	4.38	4.64

three sample sizes are more typical of individual-level survey and census data: There might be 1,000 cases in a small survey, 10,000 in a big one, and 100,000 in a census subsample, a large event-history database, or a cross-national collection of surveys. The minimum t values in Table 7 are for the most part larger than 2, suggesting that the common rule of viewing t values greater than 2 as "significant" overstates the evidence that they imply.

In the context of linear regression, equation (26) indicates that the evidence for an additional independent variable can be measured by

$$\mathrm{BIC}'_{k+1} - \mathrm{BIC}'_k = n \log\{(1 - R^2_{k-1})/(1 - R^2_k)\} + \log n, \quad (30)$$

where M_k is nested within M_{k+1}, which contains one additional variable. For there to be any evidence in favor of the new variable, the right-hand side of (30) should be negative. Thus for a BIC′ change of more than ∇BIC′, we would need to have

$$\mathrm{RED}_{k,k+1} > 1 - \exp[-(\nabla\mathrm{BIC}' + \log n)/n], \quad (31)$$

where $\mathrm{RED}_{k,k+1} = 1 - (1 - R^2_{k+1})/(1 - R^2_k)$ is the proportional reduction in residual sum of squares due to the additional variable. When R^2_k is small, then $\mathrm{RED}_{k,k+1} \approx R^2_{k+1} - R^2_k$, which is the increase in R^2 due to the additional variable, and so equation (31) becomes

$$\text{Increase in } R^2 > 1 - \exp[-(\nabla\mathrm{BIC}' + \log n)/n]. \quad (32)$$

Note that equation (32) is valid only when R^2_k is small and should be a reasonable approximation for, say, $R^2_k < .30$. The values of (31) or (32) corresponding to various grades of evidence for different sample sizes are shown in Table 8.

TABLE 8
Minimum Percent Reduction in the Residual Sum of Squares Required for Different Grades of Evidence in Favor of One Additional Variable in Linear Regression. When R^2 is small, this is roughly equal to the required increase in R^2.

Evidence	Minimum BIC Difference	n					
		30	50	100	1,000	10,000	100,000
Weak	0	10.7	7.5	4.5	0.7	.09	.012
Positive	2	16.5	11.2	6.4	0.9	.11	.014
Strong	6	26.9	18.0	10.1	1.3	.15	.018
Very Strong	10	36.0	24.3	13.6	1.7	.19	.022

TABLE 9
Approximate Two-sided *P*-Values Corresponding to Different Grades of Evidence in Favor of One Additional Parameter

Evidence	Minimum BIC Difference	n					
		30	50	100	1,000	10,000	100,000
Weak	0	.076	.053	.032	.009	.002	.0007
Positive	2	.028	.019	.010	.003	.0008	.0002
Strong	6	.005	.003	.001	.0003	.0001	.00003
Very strong	10	.001	.0005	.0001	.00004	.00001	.000004

4.4. *BIC and P-Values*

The *P*-values corresponding to the *t* statistics in Table 7 are shown in Table 9. These are rather different from the commonly used .05 and .01 cutoffs, and in most cases are smaller. For sample sizes in the 30–50 range, they are in rough agreement with conventional rules, but for larger sample sizes, much smaller *P*-values are required to imply that the data provide evidence for the effect of interest. Conventional advice has been that the significance level should decline as sample size increases, but how this should be done has not been spelled out. Table 9 provides precise guidelines for doing so, and reveals that, for large samples of the sizes that sociologists routinely work with, significance levels need to be lowered more drastically than one would perhaps have expected.

It is important to note that Table 9 is valid only for tests

involving one additional parameter (i.e., one degree of freedom). Equivalent tables could be constructed for tests with more than one degree of freedom; typically the deviation from conventional values would be even greater than where there is one degree of freedom, especially for the larger sample sizes.

In fact, the use of Bayes factors can be viewed as a precise way of implementing the advice of Neyman and Pearson (1933) that power and significance be balanced when setting the significance level, in the following sense. Suppose that half the time the null hypothesis, M_1, is true and that half the time it is false, in which case the alternative hypothesis, M_2, is true. Then the overall error rate (total of Type I and Type II errors) is minimized when the testing rule is to reject the null hypothesis whenever the Bayes factor favors the alternative—that is, whenever $B_{21} > 1$, or, approximately equivalently, when $\mathrm{BIC}_2 < \mathrm{BIC}_1$ or $\mathrm{BIC}'_2 < \mathrm{BIC}'_1$. This was shown by Jeffreys (1961, pp. 396–97), as was pointed out by Kass (1991) using more modern terminology.

It is clear from Table 9 that naive interpretations of P-values such as "$P = .001$ means that the null hypothesis is false with probability .999" are wrong. To be fair, arguments for P-values do not claim that such an interpretation is valid, but it may be a surprise that with a large enough sample ($n = 100{,}000$) $P = .001$ actually corresponds to evidence *for* the null hypothesis.

There is no real conflict between Bayes factors and significance tests: Bayes factors can be viewed as a way of setting the significance level in the test. With large samples, the appropriate level can be well below conventional levels such as .05 or .01, as Table 9 shows. However, there is a conflict between Bayes factors and significance testing at predetermined levels such as .05 or .01. There seem to be two reasons for this conflict. The first is the nature of the question posed by the P-value-based test:

> What is wrong with the likelihood ratio test? The aim of much social research is to describe the main features of selected aspects of social reality and is necessarily to some extent approximate. The LRT, in common with other significance tests, is designed to detect *any* discrepancies between model and reality. Such discrepancies do exist, by definition, al-

though if the model is satisfactory, they should be small. With a large enough sample, the LRT will find them and reject even a good model.

In the contingency table case, the LRT tests a model M_0 say, against the saturated model M_1. Assume for the moment that no other models are being considered. Rejection of M_0 then implies acceptance of M_1, which says that each cell is a special case. This does constitute a statement about the underlying social reality and may, indeed, itself be a model of interest. Rejection of M_0 does not imply that M_1 provides a better description. The point is that we should be *comparing* the models, not just looking for possibly minor discrepancies between one of them and the data.

The question to which we really want an answer can perhaps often best be expressed as follows: which model better describes the main features of social reality as reflected in the data? A closely related and more precise question is: given the data, which of M_0 and M_1 is more likely to be the true model?

The latter question can be answered by calculating the posterior odds for M_0 against M_1 (Raftery 1986*b*).

The second reason relates to the nature of the conditioning in the two procedures. A standard test rejects H_0 if equation (1) holds—that is, if the probability under H_0 of observing a value of the test statistic as extreme *or more so* is small. Thus the standard test conditions on the event $\{T \geq t(D)\}$—that is, the event that the test statistic was as extreme as the value observed, or more so. However, what *actually* happened was the event $\{T = t(D)\}$, which is less surprising under H_0 (because less extreme), and hence casts less doubt on H_0. Bayesian model selection conditions on what actually happened—namely, $\{T = t(D)\}$, suggesting the data to be less surprising under H_0 than does the standard test. Thus the Bayesian method tends to be less likely to reject a null hypothesis. Jeffreys (1980) wrote:

> I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened.

Berger and Sellke (1987) gave the following simple illustration of the distinction:[7]

> Suppose that X is measured by a weighing scale that occasionally "sticks" (to the accompaniment of a flashing light). When the scale sticks at 100 (recognizable from the flashing light) one knows only that the true value x was greater than 100. If large X casts doubt on H_0, occurrence of a "stick" at 100 should certainly be greater evidence that H_0 is false than should a true reading of $x = 100$. Thus there should be no surprise that the P-value might cause a substantial overevaluation of the evidence against H_0.

In this situation, the P-value will be the same whether or not the light is flashing, which seems counterintuitive: it is clear that there is more evidence against H_0 when the light is flashing than when it is not. In a sense, P-value-based tests *always* proceed as if the light were flashing, and that is one reason why they overestimate the evidence against H_0 in the more usual situation where the data are fully observed (or, equivalently, where the light is not flashing). By contrast, the Bayes factor for H_1 against H_0 will be greater when the light is flashing than when it is not, in agreement with intuition.

The arguments are well summarized by Berger and Sellke (1987) and Berger and Delampady (1987) and the discussants of these papers, which I recommend to the reader.

5. MODEL UNCERTAINTY AND OCCAM'S WINDOW

I now turn to the situation where there are many models, $\{M_1, \ldots, M_K\}$, and no longer just two. Suppose that Δ is a quantity of interest

[7]The quotation has been slightly paraphrased.

such as a parameter of main interest or a future observation to be predicted. Then Bayesian inference about Δ is based on its posterior distribution, which is

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|D,M_k)p(M_k|D), \tag{33}$$

by the law of total probability (Leamer 1978, p. 117). Thus the full posterior distribution of Δ is a weighted average of its posterior distributions under each of the models, where the weights are the posterior model probabilities, $p(M_k|D)$. Equation (33) provides inference about Δ that takes full account of model uncertainty.

In equation (33) the posterior model probabilities $p(M_k|D)$ are obtained by Bayes' theorem, as follows:

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{\ell=1}^{K} p(D|M_\ell)p(M_\ell)}, \tag{34}$$

which is a direct generalization of equation (5) from two models to K of them. Often all the models will be on an equal footing *a priori,* so that $p(M_1) = \ldots = p(M_K) = 1/K$. By the results in Section 4.1, approximately, $p(D|M_k) \propto \exp(-\frac{1}{2}\mathrm{BIC}_k)$ or $\exp(-\frac{1}{2}\mathrm{BIC}'_k)$. Thus

$$p(M_k|D) \approx \exp(-\tfrac{1}{2}\mathrm{BIC}_k) / \sum_{l=1}^{K} \exp(-\tfrac{1}{2}\mathrm{BIC}_\ell). \tag{35}$$

Equation (35) still holds if BIC is replaced by BIC′.

I will now consider in more detail the situation where the quantity of interest is one of the regression parameters, β_1, say. Typically some of the models specify $\beta_1 = 0$, and so the posterior probability that $\beta_1 = 0$, $\Pr[\beta_1 = 0|D]$, will be nonzero. Of particular interest is $\Pr[\beta_1 \neq 0|D]$, the posterior probability that β_1 is in the model, which is just

$$\Pr[\beta_1 \neq 0|D] = \sum_{A_1} p(M_k|D), \tag{36}$$

where $A_1 = \{M_k : k = 1, \ldots, K; \beta_1 \neq 0\}$—that is, the set of models that include β_1.

The probability that β_1 is in the model, $\Pr[\beta_1 \neq 0|D]$, can be converted to the odds scale using the relationship

$$\text{Odds} = \text{Probability} / (1 - \text{probability}),$$

and interpreted using rules of thumb such as those in Table 6. The breakpoints for weak, positive, strong, and very strong evidence are then about .50, .75, .95 and .99 on the probability scale.

Of interest also is the size of the effect, given that it is nonzero. The posterior distribution of this is

$$\begin{aligned} p(\beta_1|D, \beta_1 \neq 0) &= \sum_{A_1} p(\beta_1|D, M_k)p'(M_k|D), \\ \text{where } p'(M_k|D) &= p(M_k|D) / \Pr[\beta_1 \neq 0|D]. \end{aligned} \tag{37}$$

This can be summarized by its posterior mean and standard deviation, which may be viewed as, respectively, a Bayesian point estimator and a Bayesian analogue of the standard error. Convenient approximations to these are

$$E[\beta_1|D, \beta_1 \neq 0] \approx \sum_{A_1} \hat{\beta}_1(k)p'(M_k|D), \tag{38}$$

$$\text{SD}^2[\beta_1|D, \beta_1 \neq 0] \approx \sum_{A_1} [\text{se}_1^2(k) + \hat{\beta}_1(k)^2]p'(M_k|D) - E[\beta_1|D, \beta_1 \neq 0]^2, \tag{39}$$

where $\hat{\beta}_1(k)$ and $\text{se}_1(k)$ are respectively the MLE and standard error of β_1 under model M_k (Leamer 1978, p. 118; Raftery 1993*a*).

The main practical problem with putting this scheme into practice is that the number of models, K, may be so large that direct evaluation of the sums over all models is not feasible. For instance, in the crime example of Section 2.4, $K = 2^{15} = 32{,}768$, and so a literal implementation of the scheme would involve fitting all 32,768 regression models.

To get around this, Madigan and Raftery (1994) argued that one should exclude from the sum in (33) (a) models that are much less likely than the most likely model—say 20 times less likely, corresponding to a BIC (or BIC′) difference of 6; and (optionally) (b) models containing effects for which there is no evidence—that is, models that have more likely submodels nested within them. The models that are left are said to belong to *Occam's window,* a generalization of the famous Occam's razor, or principle of parsimony in scientific explanation. When both (a) and (b) are used, Occam's window is said to be *strict,* and when only (a) is used it is said to be *symmetric.*

Both variants of Occam's window reduce the number of models

enormously, while encompassing the essential model uncertainty present. In the crime example, there are $K = 32{,}768$ models to start with, while the symmetric Occam's window has 51, and the strict Occam's window has only 14. This is quite typical of experience to date.

A series of studies, summarized by Raftery, Madigan, and Volinsky (1995), has shown that in a range of model classes and with a variety of datasets, *taking account of model uncertainty yields better out-of-sample predictive performance than any one model that might reasonably have been selected.* This is true whether one averages across all models, or uses Occam's window in either its strict or symmetric forms. But which of these three model averaging methods is the best? The studies to date suggest that the symmetric Occam's window has predictive performance as good as that of averaging over all models, while the strict Occam's window does slightly less well predictively, but is more useful for *reporting* model uncertainty, because it involves far fewer models, and these are the most important ones. In Section 6 we report only results from the strict Occam's window.

How can we find the models in Occam's window when the initial set of models is huge? It is not feasible to proceed directly by checking each model to see whether or not it is excluded, because the number of models is too large. For the special case of linear regression, one can use the leaps and bounds algorithm of Furnival and Wilson (1974) to select a reduced set of good models, and then apply rules (a) and (b) directly to this reduced set. This is the basis for the BICREG software described in the appendix to this chapter. This has been adapted for logistic regression in the BIC.LOGIT software, which is also described in the appendix. A more general tree-based algorithm is described by Madigan and Raftery (1994); this is applicable to a wide range of model classes.

The Bayesian approach to model uncertainty was introduced by Leamer (1978). For reviews of the work since then, see Draper (1995) and Kass and Raftery (1995).

6. DIFFICULTIES RESOLVED

I now return to the practical difficulties with P-value-based tests discussed in Section 2 and describe how they are dealt with by Bayes factors, BIC, and the Bayesian approach to model uncertainty.

6.1. *Large Samples*

The BIC values for the models proposed for the large cross-national social mobility dataset of Section 2.2 are shown in Table 2. The Lipset-Zetterberg hypothesis (model 2) is indeed overwhelmingly rejected given its very large positive BIC value.[8] However, the quasi-symmetry model (model 3) is strongly preferred by BIC to the saturated model (model 4).

This agrees with the intuition of Grusky and Hauser (1984) and with the decision they made, and yet it is in dramatic conflict with the result based on *P*-values. Thus in this case BIC gives a result that is in agreement with the scientific judgment of knowledgeable investigators, while *P*-values give a result that is directly opposed to it. It is interesting to note that when Grusky and Hauser decided to ignore the *P*-value, because they felt that it clearly did not make scientific sense, they did not know about BIC and so did not have any formal statistical justification for their decision. This was the original example of BIC for log-linear models (Raftery 1986*b*). The fifth model in Table 2 is discussed below in Section 7.

6.2. *Many Candidate Independent Variables*

It was shown in Section 2.3 that when there are many candidate independent variables, statistical conclusions based on the selected model can be very misleading. They tend to identify seemingly strong relationships when, in fact, none exist. This was most strikingly illustrated by Freedman's (1983) simulation of 50 independent variables with 100 cases all consisting of pure noise unrelated to the dependent variable. In my replication of this, stepwise regression led to a highly significant and apparently satisfactory model with four independent variables (Table 3).

When the strict Occam's window was applied to these simulated data, it found five almost equiprobable models including the null model itself. When $\Pr[\beta_j \neq 0|D]$ was calculated for each variable, it was found to be zero for 44 of the 50 variables, below ½ for a further four, while for the remaining two it was 0.70 and 0.73. Even

[8]The same result holds when only the nine industrialized countries are included.

for these last two the evidence for an effect is weak on the scale of Table 6, with posterior odds of 2.3 and 2.7. Thus the conclusion from Occam's window would be that there is at most weak evidence for the inclusion of any variable, and that the null model itself is a plausible candidate. Unlike the conclusions that follow from screening methods and stepwise regression, this is not a misleading conclusion. Thus Occam's window seems to resolve the dilemma posed by Freedman's result.

It might be objected that Occam's window (and methods based on Bayes factors and BIC more generally) tends to favor parsimony to such an extent that it might find no signal even when there was one. To check whether this was so, I did two further small simulation experiments, using the same X matrix as that reported in Section 2.3. In both experiments, instead of Y being noise, Y was allowed to depend only on X_1: Y was simulated as $Y = \beta X_1 + \epsilon$, where $\epsilon \sim N(0, 1 - \beta^2)$, so that the "true" R^2 is β^2.

In the first experiment, $\beta = .45$ so that $R^2 = .20$. There Occam's window contained just one model: the correct one with X_1 only. Thus the correct conclusion was drawn by Occam's window in this case without any ambiguity or uncertainty. By contrast, the screening method described in Section 2.3 (screening out clearly nonsignificant variables from the full equation) yielded a model with 10 variables of which three were significant at the .05 level, and a P-value of 3×10^{-6}. Stepwise regression yielded a model with two variables (including X_1), both of them significant at the .05 level.

In the second experiment, $\beta = .32$, so that the true R^2 was only .10. Occam's window yielded two models with almost equal probabilities, one containing only X_1 and the other consisting of (X_1, X_{10}). Thus $\Pr[\beta_1 \neq 0|D] = 1$ and $\Pr[\beta_{10} \neq 0|D] = .52$, while $\Pr[\beta_j \neq 0|D] = 0$ for all other 48 coefficients. Thus Occam's window would lead us to conclude that X_1 certainly has an effect, that there is some very weak evidence for X_{10} having an effect, while there is no evidence that any of the other 48 variables has an effect. This is strikingly faithful to the reality, especially given the low "true" R^2 (.10), the relatively small sample size (100), and the large number of irrelevant variables (49).

By contrast, the screening method gave a model with 11 variables of which four were significant at the .05 level, while stepwise regression gave a model with two variables (including X_1) both significant at the .05 level. Once again, standard variable selection strate-

gies misleadingly detected evidence for effects of variables that were in fact not at all associated with the dependent variable.

6.3. *Model Uncertainty*

I now return to the crime example of Section 2.4, in which there was clear model uncertainty. Different variable selection methods gave quite different models. Also, in terms of the main questions of interest, different models selected gave very different estimates of β_{14}, the effect of probability of imprisonment, and also yielded different conclusions about whether X_{15}, the average time spent in state prisons, has an effect.

The Occam's window analysis of the crime data is shown in Table 10. There are 14 models, between them giving a picture of the model uncertainty in the data. Ehrlich's models do not fit well enough to be included in Occam's window, and they have BIC′ values that are far worse than the best model, by 25 and 30 points respectively. The theory on which Ehrlich's models are based would have to be very solid indeed to justify their being used as the basis for conclusions.

For X_{14}, the probability of imprisonment, the probability of an effect is high at 98 percent and the point estimate taking into account model uncertainty is -0.24. Interestingly, this is about halfway between the value from stepwise regression (-0.19) and those from the full model and the models chosen by C_p and adjusted R^2 (-0.30) in Table 5. The posterior standard deviation of β_{14} is 0.10, while for the stepwise regression model the standard error was 0.07; the difference is due to model uncertainty. The one-model standard error underestimates uncertainty, because it ignores the component due to model uncertainty.

For X_{15}, the average time spent in state prisons, the overall posterior probability that it has an effect is 0.35. Thus the data provide no evidence for this variable to have an effect, but they do not exclude this possibility either.

As for the other variables, there is very strong evidence that education and income inequality are associated with higher crime rates (each with "crime elasticities" greater than 1), positive but not strong evidence for effects of the proportions of young males and of

TABLE 10
Occam's Window Analysis of the Crime Data

# Variable	Model 1	2	3	4	5	6	7	8	9	10	11	12	13	14	Prob (%)	Post. mean	Post SD
1 % young male	•	•	•	•	•	•	•		•	•		•	•	•	94	1.40	0.50
2 South															0	—	—
3 Education	•	•	•	•	•	•	•	•	•	•	•	•	•	•	100	2.12	0.50
4 Police 1960	•	•		•	•		•	•	•			•		•	76	0.95	0.20
5 Police 1959			•			•				•	•		•		24	0.97	0.19
6 Labor part.															0	—	—
7 Sex ratio															0	—	—
8 Population						•		•			•				12	−0.08	0.04
9 Nonwhites	•	•	•	•		•		•	•	•	•	•			83	0.10	0.04
10 Unemp. 14–24															0	—	—
11 Unemp. 35–39	•	•	•		•		•						•		68	0.32	0.13
12 GDP															0	—	—
13 Inequality	•	•	•	•	•	•	•	•	•	•	•	•	•	•	100	1.33	0.32
14 Prob. prison	•	•	•	•	•	•	•	•	•	•	•	•		•	98	−0.24	0.10
15 Prison time	•			•						•					35	−0.30	0.15
R^2 (%)	84	83	82	82	80	82	80	80	80	81	79	79	78	78			
# vars.	8	7	7	7	6	7	6	6	6	7	6	6	5	5			
BIC′ (+50)	−5.9	−5.4	−4.4	−3.8	−3.6	−3.1	−2.7	−2.4	−2.4	−1.5	−1.3	−1.2	−0.9	− 0.9			
PMP (%)	24	18	11	8	8	6	5	4	4	3	2	2	2	2			

Notes: For fuller definitions of the independent variables, see Table 4.
"Prob" denotes $\Pr[\beta_j \neq 0|D]$ and is given by equation (36).
The posterior mean and SD are given by equations (38) and (39).
PMP denotes "posterior model probability" and is given by equation (35).

nonwhites, and weak evidence for an effect of unemployment among males aged 35 to 39.

The case of police expenditures is interesting. This has been measured in two successive years, and the measures are very highly correlated ($r = .993$). The data show clearly that the 1960 crime rate is associated with police expenditures, and that only one of the two measures (X_4 and X_5) is needed, but they do not say for sure which measure should be used. Each model in Occam's window contains one measure or the other, but not both. And we have $\Pr[\beta_4 \neq 0|D] + \Pr[\beta_5 \neq 0|D] = 1$, so that the data provide very strong evidence for an association with police expenditures.

The coefficient for police expenditures is positive, which may be contrary to expectations. It does indicate that increased police expenditures are not associated with lower crime rates, and hence that police expenditure is not a confounding variable for inference about the effect of X_{14} on Y—for example, at least not in the way one might expect.[9] A simple way of dealing with this is to exclude from Occam's window models with any coefficient in the wrong direction; here this would amount to excluding X_4 and X_5 and redoing the analysis.[10] Note that if the purpose of the modeling exercise is solely to *predict* crime rates (for example, in the three states not included in the data), rather than to make inference about causal mechanisms, then models with X_4 and X_5 should be included, even if the coefficients have the "wrong" sign.

There is no evidence for an effect of any of the other variables, and in the case of five of them (those for which $\Pr[\beta_j \neq 0|D] = 0$), there is evidence *against* an effect.

A more exact Bayesian analysis of these data that does not rely on the BIC′ approximation was done by Raftery, Madigan, and Hoeting (1993).

7. MODEL-BUILDING STRATEGY

One apparent difficulty with the approach outlined here is that when a parsimonious but ill-fitting model M_1 is compared with a highly

[9]One possible explanation is that increases in the crime rate lead to increased police expenditure. Time-series data would be needed to address the issue properly.

[10]This is roughly equivalent to the more sophisticated Bayesian approach of using a prior distribution for β_4 and β_5 that excludes positive values.

over-parameterized model M_2, BIC often prefers the more parsimonious model, even though it may be clearly sociologically unacceptable. When forced to choose between two unsatisfactory models, BIC tends to choose the one with fewer parameters. This has led some researchers to worry that BIC is biased in favor of parsimony over fit.

Formally speaking, this worry is unfounded, given that one ever considers only M_1 and M_2. Bayes factors are designed to choose the model that provides better out-of-sample predictions on average (Kass and Raftery 1995, sect. 3.2), and their use as a significance test minimizes the total error rate. In practice, however, when this occurs it can be an indication that neither M_1 nor M_2 is a very good model, in that M_1 may be missing an important aspect of the underlying phenomenon, while M_2 may be using too many parameters to represent it, for several of which there is no evidence.

A reasonable course of action when this happens is to search for a further model, M_3 say, which achieves most of the improvement in deviance or maximized likelihood in going from M_1 to M_2, but uses fewer parameters to do it. One way of doing this is to ask why M_2 should fit better than M_1, and then build a model that has one parameter (or so) for each reason or mechanism given. Another, complementary, approach is to inspect the residuals from M_1 to see if there is a pattern or if they can be predicted by other variables not in M_1. The resulting model, M_3, or some variant of it, may well have a better BIC value than either M_1 or M_2. Thus BIC can be used to guide an iterative model-building process.

This is well illustrated by the cross-national social mobility dataset of Sections 2.2 and 6.1. Grusky and Hauser (1984) noted that the quasi-symmetry model was preferable to the saturated model which asserts that the mobility regime in each country is different. They nevertheless searched for systematic patterns in cross-national differences between mobility regimes, explained by characteristics of the countries studied that might be expected to affect social mobility.

This led to model 5 of Table 2 above, in which the country-specific mobility parameters are allowed to vary systematically as functions of industrialization, educational participation, social democracy and inequality, with a dummy variable for Hungary. By a conventional P-value-based significance test, this model would be strongly rejected in favor of the quasi-symmetry model (and the

saturated model also), but Grusky and Hauser (1984) used it and claimed that its good fit provides evidence of systematic cross-national variation in mobility parameters. Once again, their intuitively based support of this model was (retrospectively) validated by BIC, which supports this model over the quasi-symmetry model; see Table 2.[11]

A second illustration, also from the area of social mobility, is provided by the model selection process in Hout (1988), part of which is shown in Table 11. Hout's article is about gender differences and changes over time in social mobility in the United States over the period 1972–1985. His starting point was the four-way $2 \times 3 \times 17 \times 17$ cross-classification of gender (S) $\times$ period (P) $\times$ father's occupation (O) $\times$ current occupation (D), and he used log-linear models.

Model 1 in Table 11 can be viewed as a kind of baseline model; it does not contain the $[OD]$ interaction and so would not be sociologically acceptable. Model 2 does include the $[OD]$ association but uses no fewer than $16 \times 16 = 256$ parameters to represent it. The result is a decrease in deviance that is substantial but not enough to justify the large number of parameters used to achieve it, according to BIC.

The surprising fact that BIC prefers model 1 to model 2 in Table 11 led Hout to ask how the $[OD]$ association in model 2 (which was responsible for most of the 1883-point decrease in deviance) could be more parsimoniously and interpretably represented. The answer was that the occupations of fathers and sons are associated because they have similar statuses, levels of on-the-job autonomy, and job-specific training. Using these ideas, the $[OD]$ interaction can be represented using far fewer than 256 parameters, each of which has a direct interpretation. This is achieved using Hout's own (1984) status-autonomy-training (SAT) model. The result was model 3 in Table 11, which parsimoniously represents the full four-way $[SPOD]$ interaction and has a much better BIC value than either model 1 or model 2.

[11]I have not discussed the possible presence of overdispersion in these data. Given the sample design, it is hard to see what the source of substantial overdispersion would be. In any event, if overdispersion were explicitly taken into account using standard methods (McCullagh and Nelder 1989), the deviances would be deflated and the evidence for the more parsimonious models would be stronger. Among the models considered here, the choices made would be unaffected.

TABLE 11
Fit of Models for the Four-Way Table of U.S. Mobility 1972–1985 (n = 9,227).

Model	Marginals Fitted	Deviance	d.f.	BIC
1 Table 4, model 3	$[SPO][SD]$	2653	1066	−7079
2 Table 4, model 10	$[SPO][SPD][OD]$	770	781	−6360
3 Table 5, SAT model	$[SP(SAT)]$	1167	990	− 7872

Note: O = origin occupation (17 categories); D = destination occupation (17 categories); S = gender; P = period (3 categories); (SAT) = $[OD]$ interaction parameterized using Hout's (1984) SAT model.

Source: From Hout (1988).

Thus Hout's (1988) iterative model search guided by BIC led to a model that fits better than others and is parsimonious, with each parameter being substantively interpretable. The parameter estimates (Table 5 of Hout [1988]) showed clearly how the associations between origins and destinations changed between 1972 and 1985. This clarity would have been harder to achieve with other, overparameterized, models considered.

8. DISCUSSION

In this chapter I have described the Bayesian approach to hypothesis testing, model selection, and accounting for model uncertainty. Some of the main points I have tried to argue are the following:

- Bayes factors provide a better assessment of the evidence for a hypothesis than P-values, particularly with large samples.
- Bayes factors allow the direct comparison of *nonnested* models, in a simple way.
- Bayes factors can quantify the evidence *for* a null hypothesis of interest (such as a convergence hypothesis or a theory about societal norms). They can distinguish between the situation where a null hypothesis is not rejected because there is not enough data, and that where the data provide evidence for the null hypothesis.
- BIC (or BIC′) provides a simple and accurate approximation to Bayes factors.
- When there are many candidate independent variables, standard model selection procedures are misleading and tend to find strong

evidence for effects that do not exist. By conditioning on a single model, they also ignore model uncertainty and so understate uncertainty about quantities of interest.

- Bayesian model averaging enables one to take into account model uncertainty and to avoid the difficulties with standard model selection procedures.
- The Occam's window algorithm is a manageable way to implement Bayesian model averaging, even with many models, and allows effective communication of model uncertainty.
- BIC can be used to guide an iterative model selection process.
- The methods described here can be implemented using only the output from standard statistical model-fitting software.
- Some software to implement Bayesian model averaging automatically is available.

I know of no non-Bayesian way of dealing with the model uncertainty problem. One proposal is to bootstrap the entire model-building process, including model selection. However, there is no theoretical justification for this, and Freedman, Navidi, and Peters (1988) have shown that it does not give satisfactory results. The same is true of the jackknife.

Bayesian model selection does not remove the need to check whether the models chosen fit the data. Even if many models are considered initially, they may *all* be bad! Thus diagnostic checking, residual analysis, graphical displays, and so on, all remain essential.

I have emphasized the difficulties with *P*-value-based tests in large samples, but there are difficulties also in small samples, such as arise especially in macrosociology. There, tests at a .05 level often fail to reveal any effects, which has been a source of frustration for those doing comparative and historical research (e.g., see Ragin 1987). The use of BIC corresponds to a particular sample-size-dependent choice of significance level and, as Table 9 shows, for samples sizes below about 50, that level is *greater* than .05. Thus with small samples BIC is actually *less* stringent than significance tests at a .05 level, and so BIC may provide a more satisfactory basis for the use of statistical models in comparative and historical research, as well as other areas with small samples.

BIC was introduced as a large-sample approximation to the

Bayes factor, and one may ask how large the sample has to be for it to be used.[12] That question remains to be answered, but in empirical investigations Raftery (1993*b*) found BIC to be quite accurate in examples with as few as about 40 observations. Small and unreported numerical experiments suggest it to be surprisingly accurate even for much smaller samples than that, but more research is needed on this issue. For generalized linear models, the much more accurate approximation of Raftery (1993*b*) can be used with small samples; this is implemented in the GLIB software described in the appendix to this chapter.

I have focused on the choice of independent variables in regression and related models in this chapter. However, model selection is much broader than this and also includes such modeling decisions as the coding of variables, the choice of functional forms and variable transformations, error distributions, and whether or not to remove outliers. The general framework of Bayesian model selection can be applied to these problems also. For a practical implementation of Bayesian model selection in linear regression to include the choice of independent variables, variable transformations and outlier removal, see Hoeting (1994).

What is the role of theory in all of this? Theory is essential and should be used to the greatest possible extent to define the model to be used. Indeed, the ideal situation is one in which there is no model uncertainty whatever. This ideal is sometimes approached, especially in the study of topics on which there has already been a great deal of research. Unfortunately, however, theory is often weak and vague, and does not fully specify which control variables should be included, which functional forms should be used, what the distribution of the error term is, and so on. This is often particularly the case when there has not been much previous research on the phenomenon under study. Statistical methods for model selection and accounting for model uncertainty should be used only to address issues left unresolved by theory. Bayesian model selection is not an all-purpose panacea: strong theory, clear conceptualization and careful measurement remain vital for successful social research.

[12]Bayesian model selection itself in its exact form places no restrictions on sample size, and can be used validly with even a single observation (although in that case it is unlikely to reveal much evidence for or against any model!).

APPENDIX: SOFTWARE

The BIC or BIC′ approximation can be readily calculated using the output from most standard statistical model-fitting software. All that is needed is that they return either the deviance or the LRT statistic against a null model, along with the number of parameters or the degrees of freedom.

Finding the models in Occam's window and averaging across them to account for model uncertainty can also be done using only the output from standard software, but it is much more time-consuming. I will now describe three pieces of software that help to make it more automatic.

A.1. *BICREG: Bayesian Model Selection for Linear Regression*

BICREG is an S-Plus function which can be obtained free of charge by sending the E-mail message "send bicreg from S" to the Internet address *statlib@stat.cmu.edu*. It implements the Occam's window algorithm for linear regression using the BIC′ approximation of equation (26).

For a given dependent variable and set of candidate independent variables, the software finds the models in Occam's window and their posterior probabilities, and for each independent variable it finds $\Pr[\beta_j \neq 0|D]$ and the posterior mean and standard deviation. It was used to carry out the analysis in Table 10.

It uses the leaps and bounds algorithm of Furnival and Wilson (1974) to identify a reduced set of good models. When there are more than 30 variables, it first uses backward elimination to reduce the initial set of variables to 30.

A.2. *BIC.LOGIT: Bayesian Model Selection for Logistic Regression*

BIC.LOGIT is another S-Plus function that can be obtained free of charge by sending the E-mail message "send bic.logit from S" to *statlib@stat.cmu.edu*. It is an adaptation of BICREG to the logistic regression setting and gives the same outputs.

It exploits the fact that at the MLE, logistic regression is approximately a weighted least squares problem with an adjusted dependent variable (McCullagh and Nelder 1989). To reduce the set

of models to a manageable number, it converts the logistic regression problem to the equivalent weighted least squares problem and applies a liberal version of BICREG. It then calculates BIC exactly for the remaining models, and finds those that lie in Occam's window.

A.3. *GLIB: Generalized Linear Bayesian Modeling*

GLIB is another S-Plus function that can be obtained free of charge by sending the message "send glib from S" to *statlib@stat.cmu.edu.* It does Bayesian model selection and accounting for model uncertainty for generalized linear models, notably logistic regression and log-linear models.

It differs from BICREG in two main respects, in addition to the class of models it deals with. It does not use the BIC approximation but instead carries out a more exact Bayesian analysis using a reference set of prior distributions (Raftery 1993*b*). Results are given for a range of priors. It does not yet implement Occam's window or any model search algorithm but requires the user to specify all the models to be considered. An epidemiological application was reported in detail by Raftery and Richardson (1995).

REFERENCES

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76:526–36.

Berger, James O., and Mohan Delampady. 1987. "Testing Precise Hypotheses (with Discussion)." *Statistical Science* 3:317–52.

Berger, James O., and Thomas Sellke. 1987. "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence (with Discussion)." *Journal of the American Statistical Association* 82:112–22.

Bernardo, José M., and Adrian F. M. Smith. 1994. *Bayesian Theory.* New York: Wiley.

Bishop, Yvonne M. M., Steven E. Fienberg and Paul W. Holland. 1975. *Discrete Multivariate Analysis.* Cambridge: MIT Press.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables.* New York: Wiley.

Brier, S. S., and Steven E. Fienberg. 1980. "Recent Econometric Modeling of Crime and Punishment: Support for the Deterrence Hypothesis?" *Evaluation Review* 4:147–91.

Cox, David R. 1961. "Tests of Separate Families of Hypotheses." *Proceedings of the Fourth Berkeley Symposium* 1:105–23.

———. 1962. "Further Results on Tests of Separate Families of Hypotheses." *Journal of the Royal Statistical Society (Series B)* 24:406–24.

Draper, David. 1995. "Assessment and Propagation of Model Uncertainty (with Discussion)." *Journal of the Royal Statistical Society (Series B)*, 57:45–98.

Edwards, Ward, Harold Lindman, and Leonard J. Savage. 1963. "Bayesian Statistical Inference for Psychological Research." *Psychological Review* 70:193–242.

Ehrlich, Isaac. 1973. "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation." *Journal of Political Economy* 81:521–65.

Featherman, David L., Frank L. Jones, and Robert M. Hauser. 1975. "Assumptions of Mobility Research in the United States: The Case of Occupational Status." *Social Science Research* 4:329–60.

Fenech, A., and Peter Westfall. 1988. "The Power Function of Conditional Log-Linear Model Tests." *Journal of the American Statistical Association* 83: 198–203.

Fienberg, Steven E., and William M. Mason. 1979. "Identification and Estimation of Age-Period-Cohort Effects in the Analysis of Discrete Archival Data." In *Sociological Methodology 1979*, edited by Karl F. Schuessler, 1–67. Washington: American Sociological Association.

Freedman, David A. 1983. "A Note on Screening Regression Equations." *The American Statistician* 37 (2):152–55.

Freedman, David A., W. C. Navidi, and Steven C. Peters. 1988. "On the Impact of Variable Selection in Fitting Regression Equations." In *On Model Uncertainty and its Statistical Implications*, edited by T. K. Dijkstra, pp. 1–16. Berlin: Springer-Verlag.

Furnival, G. M., and R. W. Wilson Jr. 1974. "Regression by Leaps and Bounds." *Technometrics* 16:499–511.

Grusky, David B., and Robert M. Hauser. 1983. "Comparative Social Mobility Revisited: Models of Convergence and Divergence in Sixteen Countries." Working paper, Center for Demography and Ecology, University of Wisconsin.

———. 1984. "Comparative Social Mobility Revisited: Models of Convergence and Divergence in Sixteen Countries." *American Sociological Review* 49: 19–38.

Halaby, Charles L., and David L. Weakliem. 1993. "Ownership and Authority in the Earnings Function: Nonnested Tests of Alternative Specifications." *American Sociological Review* 58:16–30.

Hazelrigg, Lawrence E., and Maurice A. Garnier. 1976. "Occupational Mobility in Industrial Societies: A Comparative Analysis of Differential Access to Occupational Ranks in Seventeen Countries." *American Sociological Review* 41:498–511.

Hoeting, Jennifer A. 1994. "Accounting for Model Uncertainty in Linear Regression." Ph.D. diss., University of Washington.

Hout, Michael. 1983. *Mobility Tables*. Beverly Hills: Sage.

———1984. "Status, Autonomy, and Training in Occupational Mobility." *American Journal of Sociology* 89:1379–409.

———. 1988. "More Universalism, Less Structural Mobility: The American Occupational Structure in the 1980s." *American Journal of Sociology* 93:1358–400.

Jeffreys, Harold. 1935. "Some Tests of Significance, Treated by the Theory of Probability." *Proceedings of the Cambridge Philosophical Society* 31:203–22.

———. 1961. *Theory of Probability,* 3rd ed. Oxford: Oxford University Press.

———. 1980. "Some General Points in Probability Theory." In *Bayesian Analysis in Econometrics and Statistics,* edited by A. Zellner, 451–54, Amsterdam: North-Holland.

Johnstone, David. 1990*a*. "Interpreting Statistical Insignificance: A Bayesian Perspective." *Psychological Reports* 66:115–21.

———. 1990*b*. "Sample Size and the Strength of Evidence." *Abacus* 26:17–35.

Kass, Robert E. 1991. "About *Theory of Probability.*" *Chance* 4:13.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors. *Journal of the American Statistical Association,* 90:377–95.

Kass, Robert E., and Larry Wasserman. 1992. "A Reference Bayesian Test for Nested Hypotheses with Large Samples." Technical Report No. 567, Department of Statistics, Carnegie Mellon University.

Kotz, Samuel, and Norman L. Johnson, eds. 1985. *Encyclopaedia of Statistical Sciences,* vol. 5. New York: Wiley.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley.

Lee, Peter M. 1989. *Bayesian Statistics: An Introduction.* Oxford: Oxford University Press.

Lipset, Seymour M., and H. L. Zetterberg. 1959. "Social Mobility in Industrial Societies." In *Social Mobility in Industrial Society* edited by S. M. Lipset and R. Bendix, 11–75. Berkeley: University of California Press.

Lye, Diane, and April Greek. 1994. "The Emerging Two-Child Norm in America." Working paper no. 95-1, Center for Studies in Demography and Ecology, University of Washington.

Madigan, David, and Adrian E. Raftery. 1994. "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window." *Journal of the American Statistical Association,* 89:1535–46.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models.* London: Chapman and Hall.

Miller, Alan J. 1984. "Selection of Subsets of Regression Variables (with Discussion)." *Journal of the Royal Statistical Society (Series A)* 147:389–425.

———. 1990. *Subset Selection in Regression.* London: Chapman and Hall.

Morrison, Denton E., and Ramon E. Henkel, eds. 1970. *The Significance Test Controversy.* Chicago: Aldine.

Neyman, Jerzy, and Egon S. Pearson. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society* A 231:289–337.

Press, S. James 1989. *Bayesian Statistics: Principles, Models and Applications.* New York: Wiley.

Raftery, Adrian E. 1986*a*. "A Note on Bayes Factors for Log-Linear Contin-

gency Table Models with Vague Prior Information." *Journal of the Royal Statistical Society (Series B)* 48:249–50.

———. 1986*b*. "Choosing Models for Cross-Classifications." *American Sociological Review* 51:145–46.

———. 1988. "Approximate Bayes Factors for Generalized Linear Models." Technical Report No. 121, Department of Statistics, University of Washington.

———. 1993*a*. "Bayesian Model Selection in Structural Equation Models." In *Testing Structural Equation Models*, edited by K. A. Bollen and J. S. Long, 163–80. Beverly Hills: Sage.

———1993*b*. "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models." Technical Report 255, Department of Statistics, University of Washington.

Raftery, Adrian E., Steven M. Lewis, and Akbar Aghajanian. 1995. "Demand or Ideation? Evidence from the Iranian Marital Fertility Decline." *Demography* 32:159–82.

Raftery, Adrian E., Steven M. Lewis, Akbar Aghajanian, and Michael J. Kahn. 1995. "Event-History Modeling of World Fertility Survey Data." *Mathematical Population Studies*, forthcoming.

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. 1993. "Model Selection and Accounting for Model Uncertainty in Linear Regression Models." Technical Report no. 262, Department of Statistics, University of Washington.

Raftery, Adrian E., David Madigan, and Chris T. Volinsky. 1995. "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with Discussion)." In *Bayesian Statistics 5*, edited by J. M. Bernardo et al. Oxford: Oxford University Press, forthcoming.

Raftery, Adrian E., and Sylvia Richardson. 1995. "Model Selection for Generalized Linear Models via GLIB, with Application to Epidemiology." In *Bayesian Biostatistics* edited by D. A. Berry and D. K. Stangl. New York: Dekker, forthcoming.

Ragin, Charles C. 1987. *The Comparative Method.* Berkeley: University of California Press.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6:461–64.

Stigler, George J. 1970. "The Optimum Enforcement of Laws." *Journal of Political Economy* 78:526–36.

Tierney, Luke, and Kadane, Joseph B. 1986. "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association* 81:82–86.

Vandaele, Walter. 1978. "Participation in Illegitimate Activities: Ehrlich Revisited." In *Deterrence and Incapacitation,* edited by A. Blumstein, J. Cohen, and D. Nagin. 270–335. Washington: National Academy of Sciences.

Weakliem, David L. 1992. "Comparing Nonnested Models for Contingency Tables." In *Sociological Methodology 1992,* edited by Peter V. Marsden, 147–78. Oxford: Blackwell Publishers.

Weisberg, Sanford. 1985. *Applied Linear Regression.* New York: Wiley.
Xie, Yu. 1992. "The Log-Multiplicative Layer Effect for Comparing Mobility Tables." *American Sociological Review* 57:380–95.
———. 1994. "The Log-Multiplicative Models for Discrete-Time, Discrete-Covariate Event-History Data." In *Sociological Methodology 1994,* edited by Peter V. Marsden, 301–40. Oxford: Blackwell Publishers.